Published in: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial), Prague, Czech Republic, 2-4 September 2015. - Stroudsburg, PA: Association for Computational Linguistics, 2015. Pp. 149–153.

A SIP of CoFee : A Sample of Interesting Productions of Conversational Feedback

Laurent Prévot[1] **Jan Gorisch**[1,2] **Roxane Bertrand**[1] **Emilien Gorène**[1] **Brigitte Bigi**[1]
[1] Aix-Marseille Université
CNRS, LPL UMR 7309
13100 Aix-en-Provence, France
[2]Nanyang Technological University
Division of Linguistics and Multilingual Studies
Singapore 637332

Abstract

Feedback utterances are among the most frequent in dialogue. Feedback is also a crucial aspect of linguistic theories that take social interaction, involving language, into account. This paper introduces the corpora and datasets of a project scrutinizing this kind of feedback utterances in French. We present the genesis of the corpora (for a total of about 16 hours of transcribed and phone force-aligned speech) involved in the project. We introduce the resulting datasets and discuss how they are being used in on-going work with focus on the form-function relationship of conversational feedback. All the corpora created and the datasets produced in the framework of this project will be made available for research purposes.

1 Introduction

Feedback utterances are the most frequent utterance type in dialogue (Stolcke et al., 2000; Misu et al., 2011). They also play a crucial role in managing the common ground of a conversation (Clark, 1996). However, perhaps due to their apparent simplicity, they have been ignored in many linguistic studies on dialogue. The main contribution to the understanding of the feedback utterance type comes from neighboring fields: (i) Conversational Analysis (CA) has shed light on turn-taking including a careful description of *response tokens*, such as "uh-huh" (Schegloff, 1982), formerly also termed *back-channels* by (ii) computational linguist Victor Yngve (Yngve, 1970)[1]; (iii) Dialogue engineers dealt with them because of their ubiquity in task-oriented dialogues (Traum, 1994); (iv) Cognitive psychologists gave them an important role in their theory of communication (Clark, 1996); (v) The most linguistic attempt to describe feedback is the work by Allwood et al. (1992) who suggest a semantic framework for it.

We take the apparent lack of sophistication of the lexical forms and structures involved in the majority of feedback utterances to be an interesting feature for a multimodal study. In our opinion, multimodal corpus studies are suffering form a combinatorial explosion that results from the simultaneous integration of complex phenomena and structures from all levels of analysis. Our aim is to use feedback as a filtering constraint on large multimodal corpora. In this way, all the dimensions will be analyzed but in a restricted way: on feedback utterances. Feedback production is known to be dependent on the discourse situation. Therefore, a second aim is to provide a model that is not domain-restricted: our objective is rather a model that is generalisable enough to be interesting from a linguistic viewpoint.

These parameters lead us to constitute a dataset that is built from four different corpora recorded in four different situations: almost free conversation (CID corpus), Map Task (MTR corpus), Face-to-Face Map Task (MTX corpus), and discussion / negotiation centered on DVD movies (DVD corpus). Since the overall goal of the project is a study of the form-function relationship of feedback utterances, the corpora are needed to create rich datasets that include extracted features from the audio, video, and their transcriptions, as well as annotated functions of the feedback utterances.

In this paper, after coming back to definitions, terminology and related work (Section 2), we present how the corpora were created (Section 3), including various stages of non-trivial post-processing, how they were pre-segmented in the gestural domain and annotated for communicative functions. We also present the different datasets (Section 4), including automatically enriched tran-

[1]See section 2 for details on the definitions and terminology.

scriptions and large feature files, how they were produced and how they can also be useful for other researchers and their studies.

2 Feedback items

Concerning the definition of the term *feedback utterance*, we follow Bunt (1994, p.27):

> "Feedback is the phenomenon that a dialogue participant provides information about his processing of the partner's previous utterances. This includes information about perceptual processing (hearing, reading), about interpretation (direct or indirect), about evaluation (agreement, disbelief, surprise,...) and about dispatch (fulfillment of a request, carrying out a command,...)."

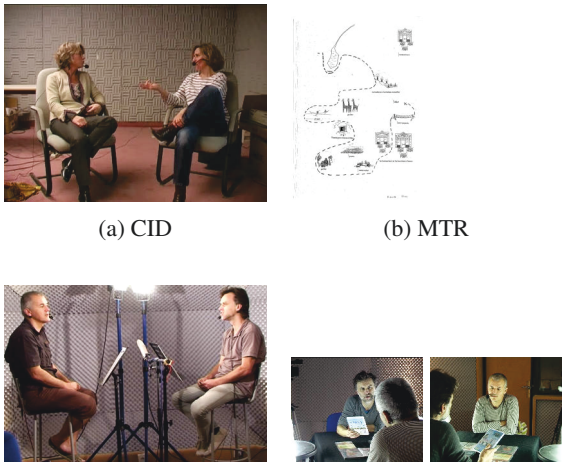
As a working definition of our class *feedback*, we could have followed Gravano et al. (2012), who selected their tokens according to the individual word transcriptions. Alternatively, Neiberg et al. (2013) performed an acoustic automatic detection of potential feedback turns, followed by a manual check and selection. Given our objective, we preferred to use perhaps more complex units that are closer to *feedback utterances*. We consider that the feedback function is expressed overwhelmingly through short utterances or fragments (Ginzburg, 2012) or in the beginning of potentially longer contributions. We therefore automatically extracted candidate feedback utterances of these two kinds. Utterances are however already sophisticated objects that would require a specific segmentation campaign. We rely on a rougher unit: the Inter-Pausal Unit (IPU). IPUs are stretches of talk situated between silent pauses of a given duration, here 200 milliseconds. In addition to these *isolated feedback IPUs*, we added sequences of feedback-related lexical items situated at the very beginning of an IPU.

3 Corpora

Our collection is composed of four different corpora: an 8 hour conversational data corpus (Bertrand et al., 2008), a 2.5 hours MapTask corpus (Bard et al., 2013), a 2.5 hours face-to-face MapTask corpus (Gorisch et al., 2014) and a 4 hours DVD negotiation corpus. All these corpora are accessible as a collection of resources through the Ortolang platform (`http://sldr.org/ortolang-000911`).

3.1 Corpus creation: Protocols, Recordings and Transcriptions

All recordings include headset microphone channels that were transcribed on IPU level and automatically aligned on word and phone level. The recording setups are illustrated in Figure 1. The first two corpora (CID and MTR) already existed before our current project, while the other two (MTX and DVD) were specifically recorded and transcribed (using SPPAS (Bigi, 2012)) for this project and are therefore explained in more detail below. CID, MTX and DVD primary are directly accessible for research purposes; MTR requires agreement from its creators.



(a) CID (b) MTR

(c) MTX (d) DVD

Figure 1: Recording setups of corpora.

CID Conversation Interaction Data (CID) includes participants having a chat about "strange things" (Bertrand et al., 2008). Each interaction took 60 minutes. Three of them were additionally recorded on video. Figure 1a illustrates the setup.

MTR The remote condition of the French MapTask corpus (MTR) (Bard et al., 2013) follows the original MapTask protocol (Anderson et al., 1991), where the role of map giver and follower change through the 8 maps per session. An example of a pair of maps is illustrated in Figure 1b. In this condition, the participants could not see each other and were therefore not recorded on video.

MTX The face-to-face condition of the French MapTask corpus (MTX) (Gorisch et al., 2014) includes additional video recordings for both partic-

ipants individually as they could see each other during the dialogue (cf. Figure 1c). Similar to the remote condition, 4 maps were "given" by one participant and "followed" by the other and vice versa. Each map took ca. 5 minutes to complete.

DVD We recruited 16 participants to take part in the recording of this corpus. The aim was to involve them in a discussion on movies, DVDs, actors, and all other topics that they may come up with during a 30 minute conversation. A set of DVD boxes (with content) were placed on a table in front of them: 4 on each side (see Figure 1d). The instructions included that each participant can take 2 of the 8 boxes home if they are on their side once the recording session is finished (as compensation for participation). Several weeks prior to the recording session, the participants were asked to fill out a short questionnaire answering four questions: what are your preferred movie genres, what are your three most preferred movies, what are your dispreferred movie genres, and what are your three most dispreferred movies. According to the answers, we paired mis-matching participants, chose 8 DVDs and placed them on the two sides in a way that maximises negotiation (who takes which DVDs home). 2 dispreferred movies or genres were placed on the own side and two preferred ones were placed on the other side.

3.2 Post-processing

Due to clocking differences in the audio and video recording devices and random image loss in the video, both signals ran out of synchronisation over time. For multimodal analyses, such desynchronisation is not acceptable. The videos of the CID corpus have been corrected by hand in order to match the audio channels. A more precise and less time-consuming procedure was developed for the newer recordings of MTX and DVD, as it is described by Gorisch and Prévot (2015). First, the audio and video files were cut in a rough manner to the approximate start time of the task, e.g. maps in the MapTask. Second, a dynamic programming approach took the audio channel of the camera and aligned it to the headset microphone mix in order to estimate the missing images for each video. Third, scripts were used to extract all images, insert images at the appropriate places and recombine the images to a film that can run synchronously with the headset microphone channels. This procedure helped to repair the videos of 2h (out of 2.5h) of the MTX corpus and the entire DVD corpus.

3.3 Gesture pre-segmentation

As our project aims to describe conversational feedback in general, the visible part of that feedback should receive sufficient attention, too. Three of the four corpora include participants' visibility and video recordings. An entire labelling of all gestures of the corpus is however impossible. Therefore, we employed two students (working on gesture for their research) to perform a pre-segmentation task. Those sections of a video that involve feedback in the domain of gestures or facial expressions were segmented using the ELAN tool in its segmentation mode (Wittenburg et al., 2006). The focus on this pass was on recall rather than precision since all the marked units will be annotated later on for precise gestures and potentially discarded if it turns out that they are not feedback.

3.4 Quantitative presentation

The content of all corpora that are included in our SIP of CoFee database, sums up to almost 17 hours of actual speech duration, with a number of 268,581 tokens in 33,378 utterances (See Table 1). This relatively large collection is used in subsequent analyses in order to quantify the form-function relationship of conversational feedback. In Table 1, the column *# Feedback* includes all (13,036) candidate feedback units (isolated IPUs and initial of an IPU). How thay have been selected is explained in Section 4. The column *# Gestures* indicates the number of pre-segmented feedback gestures. In parenthesis is the number of those gestures that co-occur with verbal feedback items. The number of gestures however should not be taken as indicator of importance of gestures in different corpora: the CID corpus has only three hours out of eight that include video-recording, while MTX misses some video files due to technical issues (see Section 3.2).

4 Datasets

This section describes how the verbal units of feedback have been selected from the transcriptions, what basic features have been extracted and what communicative functions have been (and are currently) annotated in order to form the dataset for the form-function analysis.

Corpus	# Tokens	# IPUs	actual speech duration	# Feedback	# Gestures	
CID	125,619	13,134	7h 34min	4,795	802	(516)
MTR	42,016	6,425	2h 32min	2,622	-	-
MTX	36,923	5,830	2h 33min	2,484	652	(466)
DVD	64,023	7,989	4h 12min	3,135	1,386	(668)
CoFee (all)	268,581	33,378	16h 51min	13,036	2,840	(1,650)

Table 1: Basic figures of our SIP of CoFee

Extracting units of analysis We first identified the small set of most frequent items composing feedback utterances by building the token distribution for Inter-Pausal Units (IPUs) of length 3 or less. The 10 most frequent forms are: *ouais / yeah (2781), mh (2321), d'accord / agree-right (1082), laughter (920), oui / yes (888), euh / uh (669), ok (632), ah (433), voilà / that's it-right (360)*. The next ones are *et / and (360), non / no (319), tu / you (287), alors / then (151), bon / well (150)* and then follows a series of other pronouns and determiners with frequency dropping quickly. After qualitative evaluation, we excluded *tu, et* and *alors* as they were unrelated to feedback in these short isolated IPUs. Table 2 shows the feedback tokens and the number of occurrences in each corpus. In order to count multiple sayings of a token in an IPU, such as *"oui oui"*, they appear in separate rows indicated by a plus sign (+). The category *complex* simply corresponds to any other transcription in the IPUs; it includes mainly various feedback item combinations (*ah ouais d'accord, euh ben ouais*) and repeated material from the left context. This yielded us a dataset of 13,036 utterances.

Feature extraction and function annotation In order to deepen our understanding of these feedback items, we extracted a set of form-related and contextual features. Concerning the form, aside the simplified transcription presented in Table 2, we included some features trying to describe the *complex* category (namely the presence of a given discourse marker in the unit or a repetition of the left context). Various acoustic features including duration, pitch, intensity and voice quality parameters were also extracted. Concerning contextual features, we extracted timing features within the speech environment (that provide us information about feedback timing and overlap), discourse lexical (initial and final n-grams) and acoustic (pitch, intensity, etc.) features defined in terms of properties of the previous IPU from speaker and interlocutor.

Token	CID	DVD	MTR	MTX	all
oui+	17	11	8	6	42
ouais+	141	63	26	22	252
voilà	47	41	133	105	326
ah	164	112	28	61	365
ok	5	47	132	213	397
non	109	112	103	91	415
oui	99	74	175	220	568
mh+	334	39	246	45	664
d'accord	35	83	199	366	683
mh	548	312	79	79	1,018
@	611	286	48	81	1,026
ouais	843	727	565	434	2,569
complex	1,842	1,228	880	761	4,711
Total	4,795	3,135	2,622	2,484	13,036

Table 2: Distribution of the 'simplified' transcription of IPUs.

We currently run campaigns to annotate the remaining data with feedback communicative functions (*acknowledgment, approval, answer,* etc.). Completely annotated subdatasets are used to run form-function classification experiments and correlation testing (Prévot and Gorisch, 2014).

5 Conclusion

The SIP of CoFee is ready for consumption. It is a composition of corpora of varying recording situations, including multimodality, and datasets that can be – and are currently – used for the study of one of the most basic practices in human communication, namely feedback.

Acknowledgements

This work is supported by *Agence Nationale de la Recherche* (ANR-12-JCJC-JSH2-006-01) and the *Erasmus Mundus Action 2* program *MULTI* (GA number 2010-5094-7). We would like to thank our transcribers and segmenters Aurélie Goujon, Charlotte Bouget and Léo Baiocchi.

References

J. Allwood, J. Nivre, and E. Ahlsen. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34:351–366.

E. G. Bard, C. Astésano, M. D'Imperio, A. Turk, N. Nguyen, L. Prévot, and B. Bigi. 2013. Aix MapTask: A new French resource for prosodic and discourse studies. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP)*, Aix-en-Provence, France.

R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy. 2008. Le CID-corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, 49(3):1–30.

B. Bigi. 2012. SPPAS: a tool for the phonetic segmentation of speech. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1748–1755, ISBN 978–2–9517408–7–7, Istanbul, Turkey.

H. Bunt. 1994. Context and dialogue control. *Think Quarterly*, 3(1):19–31.

H.H. Clark. 1996. *Using Language*. Cambridge: Cambridge University Press.

J. Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

J. Gorisch and L. Prévot. 2015. Audio synchronisation with a tunnel matrix for time series and dynamic programming. In *Proceedings of ICASSP 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3846–3850, Brisbane, Australia.

J. Gorisch, C. Astésano, E. Bard, B. Bigi, and L. Prévot. 2014. Aix Map Task corpus: The French multimodal corpus of task-oriented dialogue. In *Proceedings of The Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

A. Gravano, J. Hirschberg, and Š. Beňuš. 2012. Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, 38(1):1–39.

T. Misu, E. Mizukami, Y. Shiga, S. Kawamoto, H. Kawai, and S. Nakamura. 2011. Toward construction of spoken dialogue system that evokes users' spontaneous backchannels. In *Proceedings of the SIGDIAL 2011 Conference*, pages 259–265. Association for Computational Linguistics.

D. Neiberg, G. Salvi, and J. Gustafson. 2013. Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication*, 55:451–469.

L. Prévot and J. Gorisch. 2014. Crossing empirical and formal approaches for studying french feedback items. In *Proceedings of Logic and Engineering of Natural Language Semantics 11*, Tokyo, Japan.

E. A. Schegloff. 1982. Discourse as an interactional achievement: Some use of 'uh-huh' and other things that come between sentences. *Georgetown University Round Table on Languages and Linguistics, Analyzing discourse: Text and talk*, pages 71–93.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C.V. Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

D. Traum. 1994. *A computational theory of grounding in natural language conversation.* Ph.D. thesis, University of Rochester.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*. Citeseer.

V. H. Yngve. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–578.